

# GLOBAL HISTORICAL CLIMATOLOGY NETWORK (GHCN) QUALITY CONTROL OF MONTHLY TEMPERATURE DATA

THOMAS C. PETERSON<sup>a,\*</sup>, RUSSELL VOSE<sup>b</sup>, RICHARD SCHMOYER<sup>c</sup> and VYACHEVSLAV RAZUVAËV<sup>d</sup>

<sup>a</sup> National Climatic Data Center/NOAA, Asheville, NC 28801, USA

<sup>b</sup> Arizona State University, Tempe, AZ 85287, USA

<sup>c</sup> Oak Ridge National Laboratory/DOE, Oak Ridge, TN 37831, USA

<sup>d</sup> Research Institute of Hydrometeorological Information, Obninsk, Russia

Received 4 February 1997

Revised 16 March 1998

Accepted 17 March 1998

## ABSTRACT

All geophysical data bases need some form of quality assurance. Otherwise, erroneous data points may produce faulty analyses. However, simplistic quality control procedures have been known to contribute to erroneous conclusions by removing valid data points that were more extreme than the data set compilers expected. In producing version 2 of the global historical climatology network's (GHCN's) temperature data sets, a variety of quality control tests were evaluated and a specialized suite of procedures was developed. Quality control traditionally relies primarily on checks for outliers from both a time series and spatial perspective, the latter accomplished by comparisons with neighbouring stations. This traditional approach was used, and it was determined that there are many data problems that require additional tests to detect. In this paper a suite of quality control tests are justified and documented and applied to this global temperature data base, emphasizing the logic and limitations of each test. © 1998 Royal Meteorological Society.

KEY WORDS: Global historical climatology network (GHCN); quality control; temperature, monthly; homogeneity tests; outliers

## 1. INTRODUCTION

The global historical climatology network (GHCN) is a data set of monthly land surface station observations. Version 1 of GHCN was released in 1992 and contained data from approximately 6000 temperature, 7500 precipitation, and 2000 pressure stations (Vose *et al.*, 1992). GHCN is a World Meteorological Organization global baseline data set, and therefore has received considerable assistance and cooperation from many individuals, organizations and countries. The version 2 temperature data sets, with 7300 monthly mean temperature stations and 5100 monthly mean maximum and minimum temperature stations, are derived from over 30 source data sets compared to only 10 for version 1 (Peterson and Vose, 1997). This increase in data is not the only improvement for GHCN. Special homogeneity testing and adjusting techniques (Peterson and Easterling, 1994; Easterling and Peterson, 1995; Easterling *et al.*, 1996) and a new suite of quality control (QC) tests, that are applied to these data, have been developed.

In the process of creating GHCN, cautionary remarks were made that cast doubt on the quality of climate data. For example, a meteorologist working in a tropical country noticed one station had an unusually low variance. When he had an opportunity to visit that station, the observer proudly showed him his clean, white instrument shelter in a well cared for grass clearing. Unfortunately, the observer was

\* Correspondence to: National Climatic Data Center/NOAA 151 Patton Ave, Room 120, Asheville, NC 28801, USA; e-mail: tpeterso@ncdc.noaa.gov

Contract grant sponsor: DOE; Contract grant number: Interagency Agreement No. DE-AI05-900R21956

Contract grant sponsor: NOAA Climate and Global Change Data and Detection program

never sent any instruments so every day he would go up to the shelter, guess the temperature, and dutifully write it down. Another story is about a station situated next to a steep hillside. A few of meters uphill from the station was a path which students used walking to and from school. On the way home from school, boys would stop and...well, let's just say the gauge observations were greater than the actual rainfall. In the late 1800s, a European moving to Africa maintained his home country's 19th Century siting practice of placing the thermometer under the eaves on the north wall of the house, despite the fact that he was now living south of the equator. Such disheartening anecdotes about individual stations are common and highlight the importance of QC of climate data.

Historically, the identification of outliers has been the primary emphasis of QC work (Grant and Leavenworth, 1972). In putting together GHCN v2 temperature data sets (hereafter simply GHCN) it was determined that there are a wide variety of problems with climate data that are not adequately addressed by outlier analysis. Many of these problems required specialized tests to detect. The tests developed to address QC problems fall into three categories. (i) There are the tests that apply to the entire source data set. These range from evaluation of biases inherent in a given data set to checking for processing errors; (ii) this type of test looks at the station time series as a whole. Mislocated stations are the most common problem detected by this category of test; (iii) the final group of tests examines the validity of individual data points. Outlier detection is, of course, included in this testing. A flow chart of these procedures is provided in Figure 1. It has been found that the entire suite of tests is necessary for comprehensive QC of GHCN.

## 2. EVALUATION OF SOURCE DATA SETS

Not all mean monthly temperature data sets are created equally. GHCN was compiled from over 30 different source data sets. Some of these source data sets are global but the majority are from individual countries. Data was actively sought out from many people and institutions all over the world and it was a delight when a tape or floppy disk of data would arrive from Malaysia, Swaziland, or Washington, DC. Where no digital data were available, a project was started to digitize selected stations from 19th and early 20th century European colonial archives (Peterson and Griffiths, 1996).

### 2.1. Homogeneity adjusted source data

In acquiring data sets, the original observations were sought, or, in rare cases, data corrected for specific earlier errors such as an erroneous units conversion. While this goal sounds straightforward, it is sometimes difficult to obtain such observations. Observed data are often subject to inhomogeneities due to a variety of reasons, such as changes in station location, instrumentation, or time of observation. Many researchers compiling data sets spend a great deal of time and effort identifying and then adjusting the data for the various discontinuities. The data they then release tend to be their adjusted data. While the authors recognized the importance of homogeneity adjustments and have developed their own adjustment procedures for GHCN, their mission includes preserving the original observations. Because homogeneity of data is addressed separately, this valid data quality concern is not explicitly examined in the QC. Also, it is the case that no matter how good the homogeneity adjustments, there may be a better technique developed next year and therefore it is important to preserve the original data. Including homogeneity adjusted stations (e.g. data from the 1800s adjusted so they would be equivalent to data produced by 20th Century siting practices) next to stations with original data could produce spatial inhomogeneities in the early years. Therefore, the first step in GHCN QC is to exclude homogeneity adjusted source data.

### 2.2. Synoptically derived source data

Another consideration in the compilation of GHCN is the origin of the data. The most reliable monthly data come from sources that have serially complete data for every monthly report. Monthly data derived from synoptic reports transmitted over the Global Telecommunication System (GTS) are not as reliable

as CLIMAT type monthly reports. This may be due to missing data or the orders of magnitude more digitization and corresponding greater likelihood of keypunch errors. Schneider (1992) showed that synoptically derived monthly precipitation typically differs from CLIMAT monthly precipitation by 20–40%. A similar analysis performed on temperature found synoptically derived monthly temperatures differ by as much as 0.5°C from CLIMAT temperatures (M. Halpert, personal communication, 1992). Therefore, GHCN does not include monthly data that were derived from transmitted synoptic reports. While this decision does not significantly impact the quantity of historical data available for GHCN, it does decrease the quantity of near real time data available because many more stations currently report synoptically (*ca.* 8000) than send in CLIMAT reports (*ca.* 1650).

### 2.3. Consistency checks

Once the source data set was determined to meet GHCN quality requirements, a series of consistency checks was performed to make sure both this study's processing and the processing by the source data set compiler were correct. Most of these were designed to check for gross data problems and to ensure that

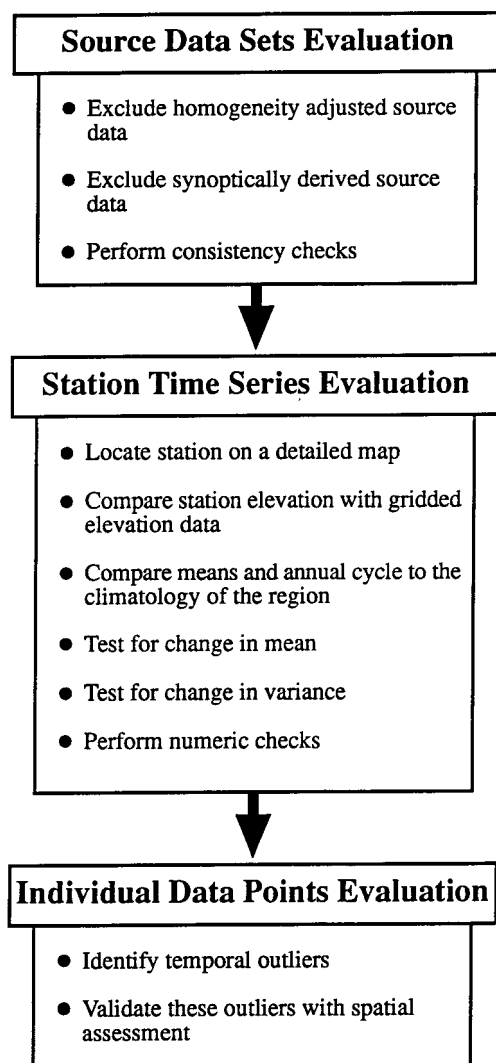


Figure 1. Flow chart of the suite of quality control tests run on the data before they are incorporated into GHCN

the data match expectations. These include making sure the authors' version of the source data set is complete and checking for special values not in the documentation. A complete list of these initial processing checks is included in Appendix A.

### 3. EVALUATION OF STATION TIME SERIES

Once the source data set passed its checks, the individual station time series were evaluated. Problems such as station mislocations and inappropriate concatenation of two different stations into a single station were detected by these tests that examined the whole time series.

#### 3.1. Climatological evaluation

The first time series check compares a station's data with the climatology of that location. A general comparison of station data means to a gridded climatology (Legates and Wilmott, 1990) can help identify questionable stations. Often the data for these stations are fine, but the stations are mislocated. One example of this was the Norwegian station of Stavanger with a digitized latitude of 38.53 instead of 58.53, erroneously placing it near the North African coast. Additional station mislocations were detected by plotting the location of each station on 1:1 000 000 Operational Navigation Charts (available from NOAA). Stations near the equator or dateline where an erroneous  $\pm$  sign would mislocate them by only a few degrees were the most common location problems detected. Since these charts had elevation contours, this plotting also indicated some erroneous station elevations. Additionally, station elevations were also compared to gridded elevation data (Row and Hastings, 1994) which can also catch gross and sometimes systematic errors in these metadata (e.g. elevation in feet instead of meters).

Besides the absolute value of the data, also evaluated was the annual cycle of the data, to see if they correspond to the climatology of the region. In areas where there is a distinct rainy season that spans the December/January annual cut off (such as some parts of southern Africa or northern Australia) it makes sense to record the data for an entire rainy season together starting with July and ending with June. Occasionally data written this way are mistakenly digitized as if they went from January to December putting the annual cycle six months out of phase. For temperature a simple relationship was used between latitude and season to provide this check and a further comparison to the gridded climatology.

#### 3.2. Testing for changes in means

It is important to identify gross discontinuities such as those caused by inappropriate concatenation of two stations into one. The Cumulative Sum (CUSUM, van Dobben de Bruyn, 1968) test looks for a change in the mean of a time series and is suited to this task. CUSUM has been used to determine the homogeneity of a station (Rhoades and Salinger, 1993).

For an observations time series  $X_1, X_2, \dots, X_n$ , the CUSUM test statistic can be written as

$$\max_{1 \leq i \leq n} (S_i - \min_{1 \leq j \leq i} S_j)$$

where

$$S_i = \sum_{j=1}^i X_j - \bar{X}$$

is the cumulative sum (Page, 1995).  $S_i$  would be high (low) for a region of the time series where observations were above (below) the mean of the whole time series. However, there are problems with CUSUM. One is that it can be sensitive to outliers. Since the intent was not to use CUSUM to detect outliers, an individual data point QC program was run (see section 4), first to remove outliers (on a preliminary basis). The second problem with CUSUM is that stations with long-term trends in temperature produce high CUSUM test statistics. To avoid this problem, each 10 year period in all the station time series was separately tested. These CUSUM scores were then normalized based on a measure of

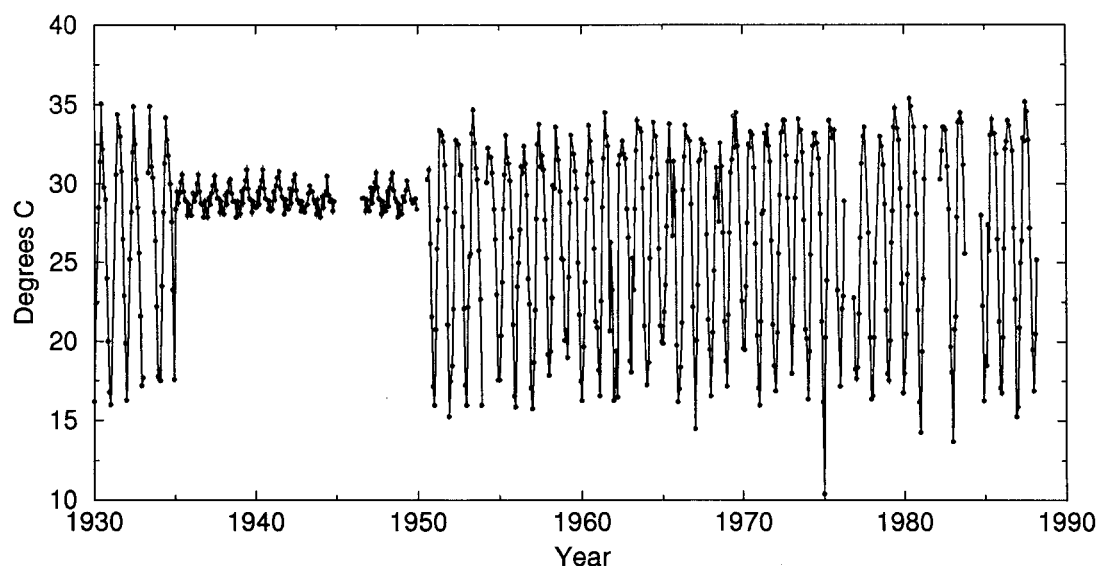


Figure 2. Mean temperature data for Bilma, Niger. For a period during the 1930s and 1940s, many stations in French West Africa reported in Kelvin. Since the mean temperature was approximate 300 K in a region with 30°C temperatures, someone 'corrected' the Kelvin temperatures by dividing by 10. Many QC checks would identify the January 'corrected' Kelvins as erroneous because they average over 10°C warmer than the mean of the other January data (28° vs. 17°), but accept most of the May data as valid because they average only 2°C warmer than the mean of the other May data (32° vs. 30°). The approach outlined in this paper identified such problems by examining the time series as a whole using the SCUSUM test we developed to identify changes in variance

variance for that station. Stations with the highest normalized CUSUM test statistic were then examined graphically and a subjective decision about the data quality was made.

### 3.3. Testing for changes in variance

The scale cumulative sum (SCUSUM) is a variant on the CUSUM that was developed to look for changes in the variance of a station time series. A good example of the type of problem that this test catches is revealed by plot of mean temperature data from the Niger station Bilma (Figure 2). In parts of French West Africa, temperatures were reported in Kelvin for a number of years (P. Jones, personal communication, 1993). With reported temperatures about 300K in a region with a mean temperature about 30°C, the values were thought to be Celsius but off by a factor of 10. The improperly 'corrected' Kelvin temperatures are clearly erroneous when shown graphically (Figure 2). However, QC that looks at individual data points might determine that all the January 'corrected' Kelvin temperatures were erroneous because they average over 10°C warmer than the mean of the other January data (28° vs. 17°), but accept most of the May data as valid because they average only 2°C warmer than the mean of the other May data (32° vs. 30°).

In SCUSUM, the cumulative sum statistic is the same but what is summed up is different:

$$S_i = \sum_{j=1}^i [(X_j - \bar{X})^2 - \sigma^2]$$

where  $\sigma^2$  is the sample variance.  $S_i$  would be large for a region of the time series where the variance is significantly larger or smaller than the variance of the whole time series. The SCUSUM was processed in the same way as was CUSUM. As with CUSUM, this test found only a few problem stations, but the problems it found were significant.

### 3.4. Numeric checks

A value of 27.2°C may seem perfectly reasonable for a given station, but when this same temperature is repeated for 4 consecutive months, its validity is questionable. Runs of numbers, both sequentially and for the same month over several years were tested for. Any station with 3 or more months identical to a precision of 0.1°C was examined in depth and a subjective decision about the quality of the station's data was made. Some stations in the tropics with very little variance and where the temperature data were digitized only to the nearest half degree, were considered valid. But for other stations some of these data points were clearly erroneous and removed.

## 4. EVALUATING INDIVIDUAL DATA POINTS

### 4.1. Temporal outliers

Just because the data values are unique and possible, based on the region's climatology, does not guarantee that they are correct, for there are many types of errors that can occur during all stages of processing (Filippov, 1968). Traditionally the easiest and most common QC technique is to test the data for outliers. Since higher latitude stations have much greater year-to-year variability than tropical stations, one must use some measure of variance to normalize the data for each station for each month prior to outlier evaluation.

**4.1.1. Measure of variance.** The most widely used measure of variance is the standard deviation (S.D.). However, using a S.D. for QC presents some problems because S.D.s are affected by the outliers themselves, so a single very large outlier can inflate the S.D. enough that one would fail to identify smaller outliers. Boyer and Levitus (1994) reduced this problem by flagging data 3–5 times the S.D. and then recalculating the S.D. without these flagged data points.

Another measure of variance is the difference between the first quartile and the fourth quartile called the interquartile range (IQR). With a large number of normally distributed data points, the IQR and the S.D. give comparable results. The IQR has been used in QC of climate data (Eischeid *et al.*, 1995) because it is very resistant to outliers. Unfortunately, because it only uses two (if the number of data points is evenly divisible by 4) or four data points (those on either side of the 25th and 75th percentile) in calculating its measure of variance, the IQR is not very robust when the time series are fairly short.

The biweight S.D. was used as a measure of variance. In the calculation of the biweight S.D.s, the median and the median of the absolute values of the deviations from this median are used to give less weight to individual observations the farther they are away from the center of the distribution (see Lanzante, 1996). Values over 5 S.D.s received 0 weight. The biweight S.D.s is robust because it uses almost all the data and the weighting factor makes it resistant to outliers. Also used were the biweight means as the center of the data distribution.

However, even using the biweight S.D., stations with 10 years of data had greater month to month differences in their variance than stations with 100 years of data. The more data points one has, the more reliable the assessment of variance. Therefore, the variance for each month was determined by using data from the month in question as well as data from the month before and the month after (after first adjusting all months to a common mean). Graphical analysis of hundreds of sample station/months revealed that this three-month approach produced excellent biweight S.D.s for the QC of temperature data.

**4.1.2. Removing long term trends.** It has been hypothesized that if a station has a significant warming trend (due to climate change or inhomogeneities), using a measure of variance that includes all the data might yield an inappropriately large variance (P. Frich in Peterson, 1994). Furthermore, under such circumstances, a warm outlier during early 'cool years' might go unnoticed. This hypothesis was tested by examining outliers from an 11-year moving median filter value rather than a mean of the whole time series. The resulting variance was much smaller. However, of the over 3 million data points evaluated,

removing a trend made almost no difference in determining which data points would be considered outliers. This indicates that the magnitude of the monthly deviations that would be considered outliers are generally much greater than observed long-term trends in the data.

*4.1.3. What threshold defines an outlier?* How warm (cold) must a month be to be considered an outlier? One common practice is using a  $3\sigma$  limit (Grant and Leavenworth, 1972; Guttman and Quayle, 1988). Alternately, Boyer and Levitus (1994) use a sigma threshold which varies from 3 to 5 depending on where the data are located. It was observed that the fraction of data that are bad is highest at the most extreme outliers and decreases as data get closer to the mean. This implies that one should be able to find a point where the number of bad data points increases sharply. To look for such a value, it is necessary to be able to determine when a datum is bad. For this analysis, it was decided that positive outliers greater than  $1.5\sigma$  could be considered bad only if all of the nearest 5 neighbouring stations had negative anomalies for that month and *vice versa* for cold outliers. This approach is based on the assumption that either (i) the 5 nearest neighbours represent the same regional climate as the station in question, or (ii) the stations where this is not true should produce a background error rate that is independent of the magnitude of the outlier, and should therefore generally stay constant over the range of outliers in the entire GHCN data set.

The results, shown in Table I, reveal that there is indeed a background error rate in our assumption that remains fairly constant at *ca.* 0.5% for data from  $1.5$  to  $2.5\sigma$ . But starting at  $2.5\sigma$ , the percentage of definitely bad data jumps up above this background rate. It is accordingly concluded that the number of errors that we can detect in the data basically starts at  $2.5$  S.D.s. Therefore, from a pure time series outlier perspective, any data point more extreme than  $2.5$  biweight  $\sigma$  was classified as suspect. But great care must be taken not to throw away good data that happens to be extreme because extreme events represent very important aspects of the climate.

## 4.2. Spatial outliers

While a data point may be extreme from a time series perspective, it may also be completely valid; i.e. it was just an exceptionally cold or warm month in that area that year. So simply flagging a data point from a time series perspective is not sufficient to judge the validity of the data point. If the climate of the region was exceptionally cold that month, nearby stations should confirm that assessment. Therefore it is necessary to integrate a spatial QC into the total assessment and use the spatial QC to determine if the 'suspect' flag should be removed from  $2.5\sigma$  outliers or not. Similar to many other QC schemes, this provides a second independent test (Smith, 1991). Unfortunately, errors can be spatially homogeneous, such as when one month a nation reports over the Global Telecommunications System that all its stations have the decimal point in the wrong place. This problem can cause false positives with any kind of spatial QC. Fortunately, such cases are fairly rare and the results are often extreme enough that data points are gross outliers. That is why data points greater than  $5\sigma$  were not accepted no matter what the spatial QC might indicate.

Table I. Percent of data at various S.D.s that could be classified as 'bad' because all 5 of the nearest neighbours were anomalies in the opposite direction

S.D.	Percentage 'bad'
1.50–1.75	0.44
1.75–2.00	0.43
2.00–2.25	0.48
2.25–2.50	0.54
2.50–2.75	0.78
2.75–3.00	1.18
3.00–4.00	1.83

Table II. Nearest neighbour comparisons

No.	Year	Mo.	C	N1	N2	N3	N4	N5	Evaluation
81	1964	1	-3.2 0	-2.7 76	-2.6 102	-1.4 106	-3.2 124	-1.9 130	Good
82	1986	5	3.7 0	3.4 57	2.8 68	1.3 68	2.5 77	2.7 90	Good
83	1987	3	-3.8 0	-3.9 89	-3.0 130	-3.9 151	-3.1 184	-3.4 192	Good
84	1968	9	3.4 0	-0.3 92	0.4 97	0.4 194	0.2 205	0.3 233	Bad
85	1967	7	-7.2 0	0.6 57	0.1 135	-0.5 186	-0.8 201	0.2 212	Bad

The last 5 of the 85 randomly selected outliers for evaluation.

C is the candidate station's temperature expressed in terms of multiples of its biweight S.D.

N1–N5 are the nearest 5 neighbouring stations, expressed in terms of their biweight S.D.

The numbers just below are the distance from the candidate station to that neighbour in km. The last column is our subjective evaluation of the data point based on a graph of the time series for that month for the candidate station's temperature and the temperature at the nearest neighbouring station.

There are many different approaches to spatial QC. One of the most ambitious is Eischeid *et al.* (1995) which uses six different methods to predict or estimate a value at any given station. These methods are: (1) the normal ratio method; (2) simple inverse distance weighting; (3) optimal interpolation; (4) multiple regression using the least absolute deviation criteria; (5) single best estimator; and (6) median of the previous five methods. The technique that created the time series most highly correlated with the data being quality controlled was used for that month's spatial QC. Different techniques might be used on the same station for different months, but, for example, all July data points for a station would be estimated by only one technique. A multiple of the interquartile range of the difference between anomalies of the observed and the estimated values was used to determine the confidence interval around the estimated value. If the anomaly of an observed value was outside the confidence interval around the anomaly of the estimated value, the data point did not pass their spatial QC.

Since Eischeid *et al.* (1995) applied their methods to GHCN version 1 data, it was of particular interest to this study. Hundreds of plots were examined of station temperature time series and flags Eischeid *et al.* (1995) generated along with the data from the station's nearest neighbour. While the subjective analysis of these plots indicated an error rate with the complex approach of Eischeid *et al.* (1995) (which was unacceptably high for our purposes), it revealed how simple spatial QC for temperature data can be and that it still works very well. Usually the single nearest neighbour could provide adequate spatial QC, though it was decided to use the 5 nearest neighbours in our technique. As part of our initial evaluation of spatial QC, we randomly selected a moderate number (85) of outliers from around the globe and took a closer look at them and their neighbours. Time series for each of these were plotted against the time series of its nearest neighbour and a subjective evaluation was made as to whether the data point was good or bad based on a single station comparison. Table II shows the last 5 of these stations, the subjective evaluation, and the magnitude of the outlier and its nearest neighbours expressed in terms of the biweight S.D.s.

Examination of all 85 sample data points revealed that almost always any one of the five nearest neighbours could give a good indication whether the station data point was valid or suspect. However, sometimes, particularly when some of the nearest stations were several hundred kilometers away, the five nearest neighbours did not provide a unanimous decision. Such a split decision can be addressed in a number of ways. One would be to go with the nearest station, but physical geographic features (e.g. a narrow mountain range between the candidate station and that nearest station) might make this station a poor choice. Another approach would be to determine which of the five stations is most highly correlated to the candidate station. However, the period of record for GHCN stations varies greatly.

Therefore, it is possible that a station 10 km away might only have that 1 y in common with the candidate station, making correlation analysis impossible.

It was decided to let any one of the five nearest neighbours provide spatial QC. This decision is unlikely to lead to very many false positives because: (i) the bulk of GHCN's data are not extreme enough to produce a 'valid' signal (e.g. 87% are less than  $1.5\sigma$ ); (ii) half the remaining data are the wrong sign to give a false positive; and (iii) as illustrated by Table II, these data are not randomly distributed. Rather, when one neighbour indicates a cold month, the other neighbours usually agree.

Because these outliers are not randomly distributed in space, the assumption of random distribution could give an indication of what threshold value for a single neighbour should be used to indicate that the outlier data point is valid. The spatial QC threshold selected is the case where the magnitude of the observed values exceeds what would be expected with a random distribution, shown in Table III. In GHCN v1 mean temperature data, 0.15% (5154) of the data points are between  $+2.75$  and  $+3.00\sigma$ . If the data were randomly distributed, 0.15% or 8 of the nearest neighbours to these 5154 data points would also be between  $+2.75$  and  $+3.00\sigma$ . Since this study checked the nearest 5 neighbours, with random distribution, 40 such matches would be expected. In the actual data, however, 970 of these 5154 data points had one of the five nearest neighbours between  $+2.75$  and  $+3.00\sigma$ , indicating that such spatial QC is not a random process.

To evaluate the appropriateness of this spatial QC based on a single neighbour, the other four neighbours were individually examined as well. If all five of the neighbours were greater than  $1\sigma$  in the right direction, that would tend to support an earlier decision to classify that data point as good, though it is certainly not proof that the data point is good (e.g. four stations at  $+1.1\sigma$  do not indicate that a  $+4.5\sigma$  outlier is valid). When applied to the entire GHCN data set, this analysis indicated that the original 'good' decision was correct 79% of the time. However, since at least one of the five nearest neighbours for some of the stations may be far away or geographically incongruent with the station in question, not passing this evaluation does not necessarily indicate an incorrect spatial QC decision. Of those data points our spatial QC indicated were bad, only 2% passed this test. While this evaluation criterion was not stringent, it does indicate that using a single station out of the five nearest neighbours for spatial QC is quite robust. Though, as with any spatial QC, the process breaks down for very remote stations, making some potentially valid outliers unlikely to be corroborated by the spatial QC. As a final step, stations with high percentage of problem data points were examined to determine the cause and appropriate remedial steps (usually removing the station from GHCN or a section of the station's time series) were taken.

## 5. SUMMARY AND CONCLUSIONS

In the final analysis, the QC decision is binary: either one uses the data point or one does not. Some data set compilers provide a series of flags based on the QC tests applied, thereby allowing, indeed forcing, the user to make the final binary QC decision based on their applications and understanding of the flags (e.g. Eischeid *et al.*, 1995). Different applications respond with different levels of resistance to bad data points,

Table III. Spatial QC threshold value

Normalized temperature ( $\sigma$ )	Threshold for 1 of the 5 nearest neighbours $\sigma$	% of suspect data reclassified as good by this threshold
4.0–5.0	1.9	56
3.0–4.0	1.8	80
2.75–3.0	1.7	86
2.5–2.75	1.6	88

To be reclassified as 'good,' one of the 5 nearest neighbours must meet or exceed this normalized value and also be an anomaly in the same direction.

but none benefit from either including bad data or excluding unusual but good data. Since, in the authors' opinion, the binary QC decision is essentially independent of the application, it should be made by those who understand the QC tests best. Therefore, rather than flagging suspect data points, we who best understand the QC tests make the final decision and remove all data points we determine as probably erroneous. However, also supplied is a companion file of these removed data points so users with specialized knowledge of historical climate events in their region of interest are able to access data points that failed the QC.

This final binary decision is the culmination of a long series of decisions that must be made all along the QC process. The QC that has been applied to GHCN temperature data is based on far more than just identification of outliers; in fact a whole suite of specialized QC tests have been developed to examine a wide variety of data problems. Indeed, some of the initial data source decisions may be the most significant decisions in the entire QC process.

Despite the problems encountered with various source data sets and individual time series, evidence was repeatedly seen, in both the digital archive and in old documents such as the 1894 *Deutsche Ueberseeische Meteorologische Beobachtungen in Deutsch-Ost-Afrika* (Peterson and Griffiths, 1997), that weather observations were generally made very meticulously. There are 4.7 million station months of temperature data in GHCN starting in 1701 and continuing to the present. This embodies the systematic observations of our environment by tens of thousands of individuals over centuries of human history. The authors feel honoured to be a part of this process and gratefully acknowledge the debt owed to the largely selfless work of individual observers. In this time of concern about our global climate, these data are becoming more important and the contributions conscientious individual weather observers made a century ago are becoming more valuable. One can obtain a copy of GHCN free of charge through the U.S. National Climatic Data Center's web site: <http://www.ncdc.noaa.gov>.

#### ACKNOWLEDGEMENTS

GHCN is jointly produced by the National Climatic Data Center/NESDIS/NOAA, the Carbon Dioxide Information Analysis Center/Oak Ridge National Laboratory/DOE, and Arizona State University. Funding for GHCN has been provided by DOE under Interagency Agreement No. DE-AI05-900R21956 and is currently being provided by NOAA Climate and Global Change Climate Change Data and Detection program. The authors would also like to thank an anonymous reviewer whose recommendations improved this paper.

#### Appendix A. INITIAL PROCESSING CHECKS APPLIED TO SOURCE DATA SETS

One very important phase of the quality assurance process involves subjecting each digital data file to extensive 'preprocessing' reviews. These reviews focus upon gross data problems (e.g. file truncations, formatting errors, unreadable records); problems that render large portions of data unusable or untrustworthy. Specific checks performed include the following:

- (i) determining if the physical characteristics of each data file (e.g. number of lines, record length) agree with supplied documentation (if any);
- (ii) determining if variable storage locations (i.e. the columns in which variables are located) and variable types (i.e. integer, real, or character) are consistent throughout each file and agree with supplied documentation (if any);
- (iii) determining the number of unique stations in each file, whether the file contains any duplicate stations and (by comparison with documentation) if any stations are missing or if any undocumented stations are present;
- (iv) determining if all date variables have reasonable values, if each file is chronologically sorted, if the period of record for each station agrees with supplied documentation (if any), and if duplicate station/date entries are present;

- (v) determining if the units of each climate variable agree with supplied documentation (if any) or determining the units if documentation is lacking; searching for specially defined values (e.g. missing value codes or trace rainfall codes) and undocumented, physically meaningless values; checking the frequency of occurrence of all possible data values;
- (vi) searching for the presence of various flag codes and whether they have documented meanings; searching for contradictions between flag codes and data values and between flag codes and other flag codes; and
- (vii) determining if the units of each metadata variable (e.g. coordinates or elevation) agree with the supplied documentation (if any) or determining the units if documentation is lacking; searching for the presence of specially defined metadata values or undocumented, meaningless values.

## REFERENCES

- Boyer, T. and Levitus, S. 1994. *Quality Control and Processing of Historical Oceanographic Temperature, Salinity, and Oxygen Data*, NOAA Technical Report NESDIS 81, Washington, D.C., 64 pp.
- Easterling, D.R., Peterson, T.C. and Karl, T.R. 1996. 'On the development and use of homogenized climate data sets', *J. Clim.*, **9**, 1429–1434.
- Easterling, D.R. and Peterson, T.C. 1995. 'A new method for detecting and adjusting for undocumented discontinuities in climatological time series', *Int. J. Climatol.*, **15**, 369–377.
- Eischeid, J., Baker, C.B., Karl, T. and Diaz, H.F. 1995. 'The quality control of long-term climatological data using objective data analysis', *J. Appl. Meteorol.*, **34**, 2787–2795.
- Filippov, V.V. 1968. *Quality Control Procedures for Meteorological Data*, World Weather Watch Planning Report No. 26, World Meteorological Organization, Geneva, 38 pp.
- Guttman, N.B. and Quayle, R.G. 1988. 'A review of cooperative temperature data validation', *J. Atmos. Ocean. Tech.*, **7**, 334–339.
- Grant, E.L. and Leavenworth, R.S. 1972. *Statistical Quality Control*, McGraw-Hill Book Company, New York, 694 pp.
- Lanzante, J.R. 1996. 'Resistant, robust and nonparametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data', *Int. J. Climatol.*, **16**, 1197–1226.
- Legates, D.R. and Wilmott, C.J. 1990. 'Mean seasonal and spatial variability in global surface air temperature', *Theor. Appl. Climatol.*, **41**, 11–21.
- Page, E.S. 1995. 'A test for a change in a parameter occurring at an unknown point', *Biometrika*, **42**, 523–527.
- Peterson, T.C. 1994. *Report of the International Workshop on Quality Control of Monthly Climate Data*, NCDC Global Climate Laboratory Monograph, 28 pp.
- Peterson, T.C. and Easterling, D.R. 1994. 'Creation of homogeneous composite climatological reference series', *Int. J. Climatol.*, **14**, 671–679.
- Peterson, T.C. and Griffiths, J.F. 1996. 'Colonial era archive data project', *Earth Syst. Monit.*, **6**, 8–16.
- Peterson, T.C. and Griffiths, J.F. 1997. 'Historical African data', *Bull. Am. Meteorol. Soc.*, **78**, 2869–2872.
- Peterson, T.C. and Vose, R.S. 1997. 'An overview of the Global Historical Climatology Network temperature data base', *Bull. Am. Meteorol. Soc.*, **78**, 2837–2849.
- Rhoades, D.A. and Salinger, M.J. 1993. 'Adjustment of temperature and rainfall records for site changes', *Int. J. Climatol.*, **13**, 899–913.
- Row, L.W. III and Hastings, D.A. 1994. *Terrain Base Worldwide Digital Terrain Data*, National Geophysical Data Center, Boulder, CO.
- Schneider, U. 1992. *The GPCP Quality-Control System for Gauge-Measured Precipitation Data*, Report of a GEWEX Workshop, Koblenz, Germany, September 14–17, WMO/TD-No. 558.
- Smith, N.R. 1991. 'Objective quality control and performance diagnostics of an oceanic subsurface thermal analysis scheme', *J. Geophys. Res.*, **96**, 3279–3287.
- van Dobben de Bruyn, C.S. 1968. *Cumulative Sum Tests: Theory and Practice*, Griffith's statistical monographs and courses, No. 24, London, Griffin.
- Vose, R.S., Schmoyer, R.L., Steurer, P.M., Peterson, T.C., Heim, R., Karl, T.R. and Eischeid, J. 1992. *The Global Historical Climatology Network: Long-Term Monthly Temperature, Precipitation, Sea Level Pressure, And Station Pressure Data*, ORNL/CDIAC-53, NDP-041. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN.